DOCUMENT RESUME

ED 350 348 TM 019 125

AUTHOR McMorris, Robert F.; Boothroyd, Roger A.
TITLE Tests that Teachers Build: An Analysis of Classroom Tests in Science and Mathematics.
PUB DATE Apr 92
NOTE 21p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 21-23, 1992).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Classroom Techniques; Competence; Grade 7; Grade 8; Interviews; Junior High Schools; Mathematics Teachers; *Mathematics Tests; Multiple Choice Tests; Questionnaires; Science Teachers; *Science Tests; *Secondary School Teachers; *Teacher Made Tests; *Test Construction; Test Content

ABSTRACT

Classroom tests developed by seventh- and eighth-grade science teachers (n=23) and mathematics teachers (n=18) were analyzed by panels of content and measurement experts. The 41 participating teachers, each of whom contributed 2 tests, completed a questionnaire, an interview, and 2 measures of competence in testing. Teachers used all major item formats in their classroom tests. Science teachers favored multiple-choice items and mathematics teachers favored computation items. Faults were found in 35 percent of completion items and 20 percent of multiple-choice items on teachers' tests. Average test quality on 6 dimensions was rated 5.0 to 5.7 on 7-point semantic differential scales. Test quality was best predicted by scores on a multiple-choice measurement competency test. The sample of classroom tests is described, evaluated, and then related to teachers' training and experience, knowledge of testing, and content of test use to learn more about this pervasive, crucial, and understudied type of testing. Three tables and one figure illustrate study findings. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
ROBERT F. McMORRIS

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED350348

Tests that Teachers Build: An Analysis of Classroom Tests in Science and Mathematics

Robert F. McMorris

Roger A. Boothroyd

State University of New York at Albany

A paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA, April, 1992.

TM019125

BEST COPY AVAILABLE

Tests that Teachers Build: An Analysis of Classroom Tests in Science and Mathematics

Robert F. McMorris, & Roger A. Boothroyd[1]

State University of New York at Albany

The typical student has probably taken more teacher-made tests than he or she has eaten fast-food hamburgers, yet we may know even less about the tests than about the hamburgers. Development of such tests is hardly a franchise operation. Few teachers are given directions or prescriptions; no organized quality control is practiced. Nevertheless, the tests remain the primary basis for a multitude of educational decisions, including grading. What are some of the characteristics of actual classroom tests, and how good are these tests judged to be? Do teachers have sufficient professional skill in test development to turn content knowledge into more than hamburger? (Food for thought?)

With the cooperation of a sample of science and mathematics teachers, we examined actual classroom tests developed by individual teachers. In addition, teachers completed two measures of competence in testing plus a questionnaire and an interview.

Research questions

What types of tests do teachers construct? (e.g., what types of items are used?)

For what purposes do teachers test?

To what extent do teachers apply sound principles of classroom testing to their own test development and usage?

Do many items contain violations of item-writing principles?

What is the judged quality of these tests?

Is test quality related to teacher characteristics? More specifically, does test quality relate to

.. teachers' measurement competence, and their ability to detect faulted items?
.. experience, number of measurement courses, measurement knowledge, or adequacy of measurement training?

---

[1]We appreciate the contributions of the 41 teachers for providing us time and tests, six raters for analyzing those tests, many graduates students for helping us increment the instruments, three reviewers for challenging comments and "continue" ratings and Drs. Robert M. Pruzek and Vicky L. Kouba for their substantial and constructive contributions to the dissertation on which this paper was based.

METHOD

Parts of this section also appear in Boothroyd, McMorris, and Pruzek (1992) and are reproduced here for the convenience of the reader.

Sample

Seventh- and eighth-grade science and mathematics teachers were selected for the study. Judging from prior research, classroom testing occurs with the greatest frequency for those grades and subjects, and such restrictions provided some degree of homogeneity.

Strong efforts were undertaken to obtain a sample that met prespecified criteria (e.g., developed their own classroom tests) yet varied in terms of the independent variables of this study (e.g., content area, experience, and type of school). Names of potential participants were obtained from a variety of sources including graduate courses at local colleges and universities, school districts, directors of teacher centers, teachers, and friends. Teachers were screened by telephone to ensure that they were either provisionally or permanently state-certified in either 7th- and 8th-grade science and/or mathematics, were teaching within their certification, had primary responsibility for constructing their own classroom tests and did not depend on an item manual accompanying the textbook. Only one teacher was excluded because of not constructing his/her own classroom test items.

The 41 participating teachers represented 25 public and private schools districts from many geographic regions in the state. No more than two teachers were selected from any one district with one exception in which four teachers were included. The districts were quite varied and included public (88%) and private (12%) schools in urban, suburban, and rural settings.

Twenty-three teachers (56%) taught 7th- and 8th-grade science while 18 taught mathematics at this level (44%). Approximately two-thirds (68%) were permanently state certified in their discipline while 13 (32%) had provisional certification. Female teachers outnumbered males by nearly a two-to-one margin (63% to 37%, respectively). The degree of teaching experience was somewhat evenly distributed, averaging 12 years but quite variable (SD = 7.2 years).

Instrumentation

Each teacher supplied the researchers with two classroom tests which he/she had developed. For each test, three judges used a rating form in responding to questions of test characteristics and quality. In addition, each teacher devoted approximately three-and-a-half hours to answering a multiple-choice test of measurement competence, identifying items containing rule violations, responding to a questionnaire, and interacting in an interview.

Test Rating Form. The rating scale was designed to describe and assess classroom tests on six dimensions that many authors of measurement textbooks suggest are important to a test's

overall test quality (e.g., Hopkins & Antes, 1985; Nitko, 1983). A preliminary version of the rating form was pilot tested with seven participants in a doctoral-level measurement course who each rated two classroom tests. The resulting form, a semantic differential, contained 39 adjective pairs.

Given that quality ratings were desired on each of the six dimensions and that some of the adjective pairs were more descriptive in nature as compared to evaluative, seven judges were asked to classify each adjective pair as either evaluative (i.e., a characteristic clearly good or bad) or descriptive. Nine items were classified as descriptive and therefore analyzed separately. The six test dimensions and the number of evaluative items per dimension: presentation/appearance (6), directions (4), length (2), content sampling (7), item construction (6), and overall quality (5).

The scale was used by two panels of three raters each, with one panel for science tests, the other for mathematics tests. Each panel consisted of a measurement specialist, a subject-matter specialist, and a person with both measurement and subject-matter expertise.

Mean ratings over items and raters were computed for each dimension and each test. Internal consistency reliabilities ranged from .60 for length to .98 for overall quality.

Measurement Competency Test (MCT). A 65-item, four-option, multiple-choice test was developed to assess teachers' knowledge of various measurement concepts specific to classroom testing. The test included items on test planning, types of items, item writing, reliability, and validity.

For the 41 teachers' responses to the final 65-item test the item difficulties were somewhat evenly distributed. Twenty items (31%) were relatively easy ($p > .7$), 23 items (35%) had moderate difficulty (.4 to .7), and 22 items (34%) proved difficult ($p < .4$). All but two items had positive item discrimination values, with 51% (33 items) having discrimination indices above .33. A more complete description of the items and the development procedures may be found in Boothroyd et al. (1992).

Item Judgment Task (IJT). Teachers reviewed 32 multiple-choice and completion items related to junior high school science and mathematics, identifying items considered "good" items and items perceived as "poor" items. Violations of recommended item writing principles (flaws) were introduced into some of the items. The 32 items were equally divided between mathematics and science, and further faceted to include an equal number of multiple-choice and completion items. Within each of the four resulting cells, 3/4 of the items (12 of 16) contained a "flaw" in item construction.

Six types of flaws were included, three in multiple-choice items and three others in completion items. Multiple-choice flaws included: (1) a cue repeated in both stem and answer, (2) the longest, most detailed option as the keyed response, and (3) options lacking homogeneity and plausibility. Flaws incorporated in completion items included: (1) blanks in either the beginning or middle of the statement, (2) nonspecific responses as possible

correct answers, and (3) omission of a nonessential word, such as a verb.

Analysis on teachers' responses to these items revealed that the greatest proportion of items (14 items/44%) were easy ($p > .7$), five items (16%) had moderate difficulty (.4 to .7), and 41% (13 items) were difficult ($p < .4$). Two items had negative discrimination values, 12 items (38%) had discrimination indices less than .1, and 12 items (38%) had discrimination levels greater than .33. A more extensive description of the IJT items, including development procedures and illustrative items, is presented in Boothroyd et al. (1992).

Interview Protocol. A 36-question interview protocol was developed as a means for providing some structure to the interviews and thus helping to ensure that consistent data were acquired for each teacher. The questions were designed to collect information on five topics: (1) the teacher's classroom testing practices and test development procedures [11 items], (2) his/her measurement training [5 items], (3) school/district policies and/or regulations specific to testing [4 items], (4) criteria the teacher used when making judgments concerning good/bad item decisions [3 items], and (5) the classroom tests submitted for review [13 items]. Given that the study was exploratory in nature, some additional questions were added for the purpose of exploring additional issues that arose during some of the initial teacher interviews.

RESULTS

Results are reported according to research questions.

What types of tests do teachers construct?

Information on teachers' tests was obtained by examining classroom tests they had developed. Of the 82 tests submitted for review (two tests per teacher), 64 (78%) were unit or chapter tests, 17 were midterm/final examinations (21%), and one (1%) was a quiz. The number of days of content the tests were designed to cover ranged from two days to 200 days. The average number of items on a unit test was 40 (SD = 32.6) while this figure was 91 items for midterms and finals (SD = 45.5). The teachers indicated that the unit/chapter tests tend not to be cumulative (i.e., do not contain material from previous tests) while midterms and finals typically cover all previously presented material. Both unit/chapter and midterm/finals are typically administered to multiple classes as indicated by an average of 67 students per unit or chapter test and 86 students per midterm or final.

In Table 1 the tests are described by item type. According to both the teachers' self-report estimates and the second author's independent analysis of their tests, computation items were the most popular for mathematics teachers and multiple-choice items for science teachers. Further, many formats were used by each set of teachers; indeed, with a more liberal definition of essays to include extended computational items, all these major types of items were used by each set of teachers.

For what purposes do teachers test?

An analysis of the teachers' responses to the interview question "Why do you test?" revealed four primary reasons in addition to a number of secondary considerations. Most frequently cited by a majority of the teachers (69%) was the response: "to assess students' mastery and understanding of the content taught in class."

The remaining three primary reasons were cited much less frequently, albeit with similar frequency to each other. Instructional reasons were cited by 33 percent of the teachers who reported that students' performance on classroom tests provide them with an indication as to which lessons were most effective and which lessons need to be retaught or remediation provided.

Grading was mentioned by 31% of the teachers. Many of these teachers did not place grading in the larger context of assessing students strengths and weaknesses but rather indicated that they had to assign course grades, and classroom tests were a means to that end.

Motivation was the fourth basic reason teachers offered for testing, and was cited by 28% of the teachers. These teachers believed that students would not do the assigned readings or seriously study the course material if tests were not given. Many of the teachers stated that their classroom tests are similar, in many respects, to other types of activities they do during class but are treated in a more formal manner by both students and teachers. As such, students perceive the tests as more important than other classroom activities, take them more seriously, and prepare for them to a greater extent.

To what extent do teachers apply sound principles of classroom testing to their own test development and usage?

Over half of the teachers (54%) indicated they generally develop some form of test plan prior to constructing a test. Although these plans are typically not formally written blueprints, at a minimum they involve a review, and frequently a listing, of the topics to be covered on the test. Most of the teachers indicated their planning process involves reviewing lesson plans, the textbook chapters, and other class material. Some teachers also reported reviewing old tests.

Once the topics are identified, most teachers begin to develop their own items or select items from other sources to assess each of the topics. Slightly over one-third of the teachers (34%) indicated that they weight topics by varying the number of items per topic. The procedures teachers described for deciding on weights for topics involved either taking into account the amount of time that was spent in class on specific topics or assessing the importance of specific material. In either case, these teachers indicated that they include more items on the test for topics which they deemed more important or for which they devoted a greater amount of class time.

Many of the teachers reported during the interviews that they use different item

formats for different types of content. These teachers indicated that the item format they used was most generally related to the cognitive level of the item. In science, for example, a number of teachers reported using alternate response (i.e., true/false, yes/no) and matching items for lower cognitive-level items, such as concept definitions or identification, while essay items were used to assess higher cognitive levels such as synthesis. Some of the teachers also distinguished between item formats requiring recognition (e.g., matching, alternate response, multiple choice) and those item formats necessitating recall (e.g., completion, short answer). Few teachers, however, indicated how they "balance" their classroom tests with respect to the issue of cognitive level.

Do many items contain violations of item-writing principles?

A sample of approximately 350 multiple-choice and completion items submitted by the teachers was examined, with flaws detected in 35% of the completion and 20% of the multiple-choice items. Most frequently observed problems in the completion items were blanks in the beginning or middle of a statement (25% of all completion items) and the request for a nonspecific response (14%). Nonhomogeneity of response options was present in seven percent of all multiple-choice items reviewed, with the same percentage having the longest option as the key. Cues were discovered in five percent of the items. Other flaws (in 6% of the items) included window dressing, no question presented in the stem, and spelling errors.

What was the judged quality of these tests?

Each test was rated by a three-judge panel using semantic-differential items. Panelists assigned above-average ratings on all six dimensions, judging appearance the highest (mean = 5.8 on a 7-point scale) and test length the lowest (mean = 5.0) (see Table 2). Overall quality was rated 5.4 on the average. Raters perceived the greatest variation among the tests in terms of appearance and in adequacy of directions (SDs = 1.3), and the least variability in item construction and adequacy of content sampling (SDs = .8).

Is test quality related to teacher characteristics?

An analysis of the ratings first revealed differences in ratings between mathematics and science tests on several dimensions, most importantly, overall quality. Given the two panels, this difference was confounded by different raters across subject areas, so these ratings were regressed by subject area and the residual used as the dependent variable for a regression analysis. The best predictor of the resulting quality-of-test variable was the score on the Measurement Competency Test (r = .37).

Even with the confounding of raters and subject matter, the overall classroom test quality is related to indices of measurement competency. One approach to these relationships is to dichotomize the group on ratings, on the Measurement Competency Test (MCT) and on the Item Judgment Task (IJT). For the 21 teachers in the bottom half on rated test quality, 11 were in the bottom half of the group on both the MCT and IJT; only 4 were in the top half on both predictors. For 20 teachers in the top half on rated test

quality, 10 were in the top half on both predictors; 2 were in the bottom half on both. These relationships are detailed in Figure 1.

These two extreme groups differ in teaching experience and measurement background, as noted in Table 3. The high group is the somewhat more experienced group. For each of the three measurement variables, the high-group mean exceeds the low-group mean by more than half a standard deviation.

DISCUSSION

Guttman (1970) expressed in a classic cartoon the imbalance of research emphasis on test design vs. test analysis. Similarly, study of classroom tests and their developers has lagged behind study of standardized, published measures. Classroom testing is the basis for such a variety of decisions involving instruction, grading, and other uses, yet as professionals we know little about the qualities and characteristics of such tests. We have done little to describe, let alone evaluate, these evaluative devices.

Every day, the number of tests taken in schools, and the number and type of decisions based on information from those tests, could perhaps best be described graphically by an astronomy professor from Cornell. And if we include the other types of assessment information used by teachers and students (see, e.g., Airasian, 1991; Stiggins, Conklin, & Bridgeford, 1986), the amount of information, the number of decisions, and the impact of those decisions becomes virtually incomprehensible. Especially given that teachers' training in formal testing is so limited, and their training in informal assessment is even more limited, we are concerned about 1) the quality of the measures, 2) the ability of the teaching professionals to provide professional interpretations of information and appropriate decisions using that information, and 3) our own ability and resolve to formulate and respond to educationally important questions.

Item types used by the science teachers in our study agree with item types found in junior high science tests by Fleming and Chambers (1983, p.33). Rank-order of occurance is the same across studies: multiple choice was most popular, followed by matching, short answer/completion, true false, and essay. For teachers more generally, however, Fleming and Chambers found the short answer/completion format most popular and matching a distant second.

For our sample, 20% of the multiple-choice items contained faults. Similarly, in the Oescher and Kirby (1990) study, "Of the 18 tests containing multiple choice items, 17 were judged to have flaws in more than 20% of these items" (p. 13). Carter (1986) also found faults in teacher-made tests. Of the tests Carter reviewed, 78% strongly favored the key in C, 86% had at least one item with a longer correct answer, 47% contained at least one stem cue, and 58% contained at least one grammatical clue.

But what are the impacts of item faults on teacher-made tests? Certainly items may be made easier by faults (Dunn & Goldstein, 1959; McMorris, Brown, Snyder, & Pruzek,

1972; Haladyna & Downing, 1989a; 1989b). Tests containing item faults are inconsistent with Nitko's (1983) principle that "test items should elicit only the behaviors which the test developer desires to observe." (p. 141) We would expect faulted items to introduce extraneous variance; such variance would, in turn, reduce somewhat the validity of descriptions and decisions based on the test.

Other, more subtle impacts are also possible. Students judge tests and their developers. Do you expect them to respect a bogus test or an incompetent test developer? How many times did your attitude about a teacher or professor change as a result of taking your first test in a course? To illustrate, how do you felt about a author who make grammatical errors? And on how many other dimensions would you as a student have been able to describe and discuss a teacher's test? Would you have considered easiness, content balance, and understanding or application vs. pedestrian knowledge? The teacher communicates so much with a test. Student attitude toward the course, the instructor, and the subject must be affected by that test and its interpretation.

> Classroom evaluation affects student in many ways. For instance, it guides their judgment of what is important to learn, affects their motivation,and timing of personal study (e.g., spaced practice), consolidates learning, and affects the development of enduring learning strategies and skills. It appears to be one of the most potent forces influencing education. (Crooks, 1988) (p. 467)

The impacts of a test's characteristics and quality, then, are not just in producing appropriate or extraneous variance on the measure itself. The impacts also include student attitudes and perceptions which affect what they bring to the next encounter of an evaluation kind.

One disheartening, anecdotal index of teacher frustration and student achievement levels came from the teacher interviews in this study. Some teachers admitted they intentionally included clues in items so some weaker students could answer some items correctly. Admittedly, if done with a sense of humor on an informal "test" that is essentially intended for review, there may easily be some positive benefit. If done when a less contaminated measure is desired, the extraneous variance may be expensive. At a minimum, intentional use of clues can be investigated in further studies.

Additional samples of teachers would provide appropriate replication. We would recommend including outcome measures assessing characteristics/quality of teacher-made tests and independent measures for measurement competency, measurement training, experience, etc. Extensions to our instruments could better specify knowledge of teachers' ability and practice in grading, reporting/communicating, sizing up, instructional pacing, and performance testing. Understanding how item characteristics and score distributions should follow from type of objective could also be tested (see Terwilliger, 1989).

An outcome of our profession's lack of emphasis on classroom assessment may be to allow standardized testing to win by default. As noted by Stiggins et al. (1986), laypersons and policymakers maintain that schooling outcomes are measured best and fairest by standardized paper and pencil tests, which severely restricts the variety of outcomes used for accountability. Similarly, research on teaching has also depended excessively on

standardized tests to represent school achievement. Such tests are not constructed to be maximally sensitive to instruction (Hanson, McMorris, & Bailey, 1986; Mehrens & Phillips, 1987). Issues concerning and techniques for assessing fit between test and curriculum are reviewed by Crocker, Miller, and Franks (1989).

Relationships of published achievement tests with instruction are being examinied in more sophisticated ways, and additional research is needed. Such investigations will likely have applicability to local districts and enhance the assessment of student achievement. Teachers, however, develop virtually all the achievement measures on which instructional decisions are based. The current emphasis on studying teachers' testing and assessing is reassuring.

REFERENCES

Airasian, P. W. (1991). Perspectives on measurement instruction. Educational Measurement: Issues and Practice, 10(1), 13-16,26.

Boothroyd, R. A. (1990). Variables related to the characteristics and quality of classroom tests: An exploratory study with seventh and eighth grade science and mathematics teachers. (Doctoral Dissertation, The University at Albany, 1990) Dissertation Abstracts International, 51/07A, 2355.

Boothroyd, R. A., McMorris, R. F., & Pruzek, R. M. (1992, April). What do teachers know about measurement and how did they find out? Paper presented at the annual conference of the National Council on Measurement in Education. San Francisco, CA.

Carter, K. (1986). Test-wiseness for teachers and students. Educational Measurement: Issues and Practice, 5(4), 20-23.

Crocker, L. M., Miller, M. D., & Franks, E. A. (1989). Quantitative methods for assessing fit between test and curriculum. Applied Measurement in Education, 2, 179-194.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. Review of Educational Research, 58, 438-481.

Dunn, T. F., & Goldstein, L. G. (1959). Test difficulty, validity, and reliability as a function of selected multiple-choice item construction principles. Educational and Psychological Measurement, 19, 171-179.

Fleming, M., & Chambers, B. (1983). Teacher-made tests: Window on the classroom. In W. E. Hathaway (Ed.), Testing in the schools (pp. 29-38). New Directions for Testing and Measurement, No. 19. San Francisco: Jossey-Bass.

Guttman, L. (1970). Interpretation of test design and analysis. In Proceedings of the 1969 invitational conference on testing problems (pp. 53-65). Princeton, NJ: Educational Testing Service.

Haladyna, T. M., & Downing, S. M. (1989a). A taxonomy of multiple-choice item writing rules. Applied Measurement in Education, 2, 37-50.

Haladyna, T. M., & Downing, S. M. (1989b). Validity of a taxonomy of multiple-choice item writing rules. Applied Measurement in Education, 2, 51-78.

Hanson, R. A., McMorris, R. F., & Bailey, J. D. (1986). Differences in instructional sensitivity between item formats and between achievement test items. Journal of Educational Measurement, 23, 1-12.

Hopkins, C. D., & Antes, R. L. (1985). Classroom measurement and evaluation. (2nd ed.). Itasca, IL: F. E. Peacock Publishers.

McMorris, R. F., Brown, J. A., Snyder, G. W., & Pruzek, R. M. (1972). Effects of violating item construction principles. Journal of Educational Measurement, 9, 278-295.

Mehrens, W. A., & Phillips, S. E. (1987). Sensitivity of item difficulties to curricular validity. Journal of Educational Measurement, 24, 357-370.

Nitko, A. J. (1983). Educational tests and measurement: An introduction. New York: Harcourt Brace Jovanovich.

Oescher, J., & Kirby, P. C. (1990, April). Assessing teacher-made tests in secondary math and science classrooms. Paper presented at the annual conference of the National Council on Measurement in Education, Boston, MA. (ERIC Document Reproduction Service No. ED 322 169).

Stiggins, R. J., Conklin, N. F., & Bridgeford, N. J. (1986). Classroom assessment: A key to effective education. Educational Measurement: Issues and Practice, 5(2), 5-17.

Terwilliger, J. S. (1989). Classroom standard setting and grading practices. Educational Measurement: Issues and Practice, 8(2), 15-19.

Table 1

Distribution of Classroom Test Items by Subject and Item Format

Item Type	Percentage from Teacher Self-reports			Percentage from Examination of Classroom Tests[a]		
	Mathematics (n=18)	Science (n=23)	Total (n=41)	Mathematics (n=36)	Science (n=46)	Total (n=82)
1. Multiple Choice	10	45	30	14	47	33
2. Alternate Response	4	12	8	3	12	8
3. Matching	3	18	11	3	18	11
4. Short Answer	9	8	8	13	12	13
5. Completion	10	5	7	14	5	9
6. Essay	0	7	4	0	3	2
7. Computation	60	2	27	41	1	19
8. Other	5	3	4	10	3	6

[a]Each teacher submitted two classroom tests for review.

Table 2

Reviewers' Ratings by Subject and Test Domain

Dimension	Number of Items	Mathematics (n=36)		Science (n=46)		Total (n=82)	
		Mean[a]	SD	Mean[a]	SD	Mean[a]	SD
1. Appearance	6	5.62	1.22	5.87	1.36	5.76	1.30
2. Directions	4	5.15	1.20	5.33	1.43	5.25	1.33
3. Length	2	4.53	.99	5.38	1.17	5.01	1.17
4. Content Sampling	7	5.22	.80	5.31	.89	5.27	.85
5. Item Construction	7	5.61	.44	5.11	.95	5.33	.80
6. Overall Quality	5	5.04	1.00	5.71	.89	5.41	.99
Composite	31	5.32	.63	5.41	.79	5.37	.72

[a] Ratings were obtained using a 7-point semantic differential format and were averaged for three-judge panels.

Table 3

Teacher Profiles Based on Their Performance on the MCT, IJT, and Ratings of Classroom Test Quality

Teacher Profile Variable	Scores on the MCT, IJT, and Ratings of Classroom Test Quality			
	Below the Medians (n=11)		Above the Medians (n=10)	
	Mean	Standard Deviation	Mean	Standard Deviation
Years of Teaching Experience	10.82	8.99	13.00	9.14
Number of Measurement Courses	.27	.47	1.30	1.34
Self-report Measurement Knowledge[a]	1.45	.82	2.00	.94
Adequacy of Measurement Training[b]	2.09	.94	3.10	1.10

Note. High and low categorizations were based on independent median splits performed on the MCT, IJT, and Ratings of Classroom Test Quality.

[a]Scale: 1 = Poor; 2 = Fair; 3 = Good; 4 = Very Good; 5 = Excellent

[b]Scale: 1 = Strongly Disagree; 2 = Disagree; 3 = Uncertain; 4 = Agree; 5 = Strongly Agree

Figure 1

Rated Test Quality Related to IJT and MCT